Ranking and Filtering the Selected Attributes for Intrusion Detection System

Phyu Thi Htun and Kyaw Thet Khaing

***Abstract*—**Many researchers have been focused on improving the performance, especially in accuracy in many Intrusion Detection System (IDSs). But, at the same time, they also realized that they need to detect in time as fast as they can. Therefore, the time consuming to match with a lot of the training data with testing or real-time connection data is becomes the bottleneck of the IDSs. Feature Selection with the ranked features from the training data and also from testing data and filtering the other unimportant features, play a vital role for the IDSs for detection and preventing to intrusions or attacks to the network. The ranking and filtering techniques, Information Gain in Random Forest (RF), will rank the weighted attribute and eliminate unimportant attributes for the detection of the attacks which trained in KDD's training datasets. The results will come out by matching using with pattern recognition algorithm, K-Nearest Neighbors (KNN) for detection the known and unknown attacks. The experimental results can compare with the proposed combine method and the results of the other machine learning algorithms.

***Keywords*—**attribute selection, intrusion detection, Random Forest, K-Nearest Neighbors, KDD.

I. Introduction

An intrusion detection system is an important component of the network architecture in an organization which attempts to protect the network computers against various kinds of possible attacks. Computer systems are exposed to increasing number of such security threats. To overcome these threats, a network intrusion detection system has to adopt the network security policies to detect and react against these threats as quickly as possible. Intrusion detection techniques fall under two categories i.e., misuse detection and anomaly detection.

In misuse detection, the IDS analyzes the information it gathers and compares it to large databases of attack signatures, while in anomaly detection, any deviation from the established profiles of normal activities is treated as an attack. Anomaly detection can detect novel attacks but has a high false positive rate, whereas a misuse detection system cannot detect new attacks. Many intrusion detection techniques today are rule-based [17], which has an intrinsic limitation of low detection rate for new attacks. Therefore, to overcome the limitations of the rule-based network intrusion detection techniques, several data mining techniques have been employed to find models that are better understandable by the data owner [18],[19]. There is a wide variety of network traffic, and the data required for detecting network intrusion is composed of various features. All the features in the network traffic are not necessarily required for intrusion detection. So we need to extract those features which have higher intrusion detection tendency. For that reason, different statistical techniques have been used to reduce the feature space.

There are four main categories of attacks found in the literature, namely: 1) DoS (denial-of-service) attacks, 2) R2L (Remote to Local) attacks, 3) U2R (User to Root) attacks, and 4) PROBE attacks. For establishing a connection, an attacker may follow the same steps e.g., establishing a connection from source IP to the target IP and sending data to the attack target [6]. In KDD99Cup dataset [7] different attacks have different connections, as some of the attacks have few network connections such as U2R and R2L whereas others may have hundreds of network connections such as DoS and Probe. There are different feature values for normal and attack connections in the packet header, and the packet contents can be used as signatures for intrusion detection.

In this paper we propose and implement a hybrid classifier based on Random Forests (RDF) and k-Nearest Neighbor (KNN) algorithm for the classification of DoS attacks in a network. Random Forests is an ensemble classification and regression approach which is unbeatable in accuracy among current data mining algorithms. Random forests algorithm has also been used in applications like prediction [20], probability estimation [4], and pattern analysis. After reducing the features we need to classify the records between other records and DOS attacks. We further optimize the selected features with the help of RF algorithm and compare our results with several different classification techniques. We evaluate our proposed approach on KDD'99Cup dataset. Experimental results show that by using the proposed approach, average detection rate is increased and at the same time average false positive rate is also decreased when compared with other algorithms.

The rest of the paper is organized as follows. Section 2 presents the related research using corresponding machine learning Algorithms. In section 3 described the KDD 99 CUP intrusion detection datasets which including DoS attacks data.

Phyu Thi Htun is study at the Ph.D programs of the Faculty of Information and Communication Technology, University of Technology(Yatanarpon Cyber City) Myanmar and doing the research in network security.(e-mail: ms.phyuthihtun@gmail.com)

Kyaw Thet Khaing was now with the Department of Computer Hardware Technology and serving the project of the network security of UCSY (e-mail: kyawthetkhaing.ucsy@gmail.com).

http://dx.doi.org/10.15242/IIE.E0314114

Section 4 introduces about the proposed system for DoS detection. The usefulness of Random Forest and k-NN, presented in Section 4, Section 5 describes the experimental results obtained, using the Random Forest for feature selection and examined with k-NN for performance evaluation analysis. Section 6 explains the conclusion and further extension to our research by using out coming results.

II. Related works

Many researchers tried to increase the performance of detection attacks by using many machine learning algorithms. Reference [3] have presented a comparative study between five data mining algorithms (ID3, C4.5, Random Forest, Multilayer Perceptron, Naive Bayes, k-Nearest Neighbor) applied to the classification of network intrusions. The Algorithms that have shown the highest prediction rate are the Random Forest and Naive Bayes Algorithms. ID3 is a supervised algorithm developed by [4] whose purpose is to build decision trees from a data set. Decision trees are very efficient as they classify new cases from the training data and test data to properly assess the quality of the tree constructed. The decision tree is built recursively. The ID3 calculates, among the remaining attributes, the ones which will generate the most information (information gain), which will classify examples of any level of the decision tree.

The literature suggests that hybrid or assembling multiple classifiers can improve the accuracy of a detection [5][6]. According to [5], an important advantage for combining redundant and complementary classifiers is to increase robustness, accuracy and better overall generalization. Reference [8] demonstrated the use of ensemble classifiers gave the best accuracy for each category of attack patterns. Ensemble methods aim at improving the predictive performance of a given statistical learning technique.

The most related work is done by [5]. They use Random Forests Algorithm [7] over rule-based NIDSs. They applied Random Forests Algorithm for only misuse detection. Misuse detection discovers attacks based on the patterns extracted from known intrusions.

Thus, novel attacks can't be detected in this network intrusion detection system. Surveys of the various data mining techniques have been proposed towards the enhancement of IDSs. Reference [9] shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers.

III. Units

Since 1999, KDD'99 [13] has been the most widely used data set for the evaluation of anomaly detection methods. This data set is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories:

(1) Denial of Service Attack (DoS)
(2) User to Root Attack (U2R)
(3) Remote to Local Attack (R2L) and
(4) Probing Attack

Table I showed the four categories and their corresponding attacks on each category. It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic.

Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants.

This testing datasets contains more examples of attacks than normal connections and the attack types are not represented equally. In this table described the attacks types in two, known and unknown. In the testing datasets, it will contains the attack which are not included in training datasets are called unknown attacks and described as the bold letters.(such as . apache2 in DoS attack type).

TABLE I
Classifying attacks on KDD Datasets

Type	Attack Name
Probing	Port-sweep, IP-sweep, Nmap, Satan , **Saint, Mscan**
Denial of Service (DoS)	Neptune, Smurf, Pod, Teardrop, Land, Back, **Apache2, Udpstorm, Processtable, Mailbomb**
User to Root (U2R)	Buffer-overflow, Load-module, Perl, Rootkit, spy, **Xterm, Ps, Http-tunnel, Sqlattack, Worm, Snmp-guess**
Remote to Local (R2L)	Guess_password, Ftp_write, Imap, Phf, Multihop, Warezmaster,Warezclient, **Snmp-getattack, Named,Xlock, Xsnoop, Send-mail**

For the relevant and redundant of KDD datasets, the NSL-KDD datasets modified those data sets and shared 4 dataset file, Train+, Train+_20Percent,Test+ and Test-21. The first two files represent for training datasets and contain the general attacks. The rest two files represent for testing datasets and contain not only general attacks but also the unknown (novel) attacks which are described with bold in Table 1. The connection for each attack type is shown in Table 2.

TABLE II
Number of connections in each attack type

Datasets	Normal	DoS	U2R	R2L	Probe	Total
Train+	67343	45927	993	54	11656	125973
Train+20 Percent	13449	9234	206	12	2289	25190
Test+	9711	7458	2421	533	2421	22544
Test-21	2152	4342	2421	533	2402	11850

IV. Experimental Results

Since 1999, KDD'99 [13] has been the most Complex relationships exist between the features, which are practically impossible for humans to discover. IDS must therefore reduce the amount of data to be processed. This is extremely important if real-time detection is desired. Reduction can occur in one of several ways.Finally, some data sources can be eliminated using feature selection

A. Data Filtering

The purpose of data filtering is to reduce the amount of data directly processed by the IDS. Some data may not be useful to the IDS and thus can be eliminated before processing. This has the advantage of decreasing storage requirements, reducing processing time and improving the detection rate (as data irrelevant to intrusion detection are discarded). However, filtering may throw out useful data, and so must be done carefully [10].

B. Feature Selection

The purpose Feature selection (also known as subset selection or variable selection) is a process commonly employed in machine learning to solve the high dimensionality problem. It selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more concise data representation.

The benefits of feature selection are multi-fold. First, feature selection greatly saves the running time of a learning process by removing irrelevant and redundant features. Second, without the interference of irrelevant, redundant and noisy features, learning algorithms can focus on most important aspects of data and build simpler but more accurate data models. Therefore, the classification performance is improved. Third, feature selection can help us build a simpler and more general model and get a better insight into the underlying concept of the task [11],[12].

C. Information Gain

In this method, the important features are calculated over multiple RDF iterations, the least important features being removed after each. The objective of using Random Forest is to reduce the impurity or uncertainty in data as much as possible .A subset of data is pure if all instances belong to the same class. The heuristic is to choose the attribute with the maximum Information Gain or Gain Ratio based on information theory.

Entropy is a measure of the uncertainty associated with a random variable. Given a set of examples D is possible to compute the original entropy of the dataset such as:

$$H[D] = -\sum_{j=1}^{|C|} P(c_j)\log_2 P(c_j)$$

where C is the set of desired class.

If we make attribute Ai, with v values, the root of the current tree, this will partition D into v subsets D_1,D_2….D_j . The expected entropy if A_i is used as the current root.

$$H_{A_i}[D] = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} H[D_j]$$

Information gained by selecting attribute A_i to branch or to partition the data is given by the difference of prior entropy and the entropy of selected branch.

$$gain(D, A_i) = H[D] - H_{A_i}[D]$$

We can choose the attribute with the highest gain to branch/split the current tree numbered.

Number footnotes separately in superscripts (Insert | Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

Please note that the references at the end of this document are in the preferred referencing style. Give all authors' names; do not use "*et al*." unless there are six authors or more. Use a space after authors' initials. Papers that have not been published should be cited as "unpublished" [4]. Papers that have been submitted for publication should be cited as "submitted for publication" [5]. Papers that have been accepted for publication, but not yet specified for an issue should be cited as "to be published" [6]. Please give affiliations and addresses for private communications [7].

Capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [8].

D. Proposed System

There are two phase in my proposed Model. This system is process of identifying the abnormal and normal instances in first and then identifying again to classify attacks types of those abnormal instances. The first phase is the training phase and intends to reduce the irrelevant features. After loading the training dataset, features are reduced by Random Forest Algorithm with information Gain. Using those features, the system will classify the normal and abnormal data in first step.

If normal, system keeps that connection in normal database. And then continuing classifies the abnormal connections in four types of attacks and keeps in attack database.

Next phase is detection phase. It will input the testing data set and mapping the data sets as the remaining features of the training data set. As the training phase, will detect normal and four attacks type and display an alert if a connection is a one of the attacks. This system is shown in Fig. 1.

We summarize the experimental results to detect attacks for intrusion detection with over the NSL-KDD datasets. Experimental results are presented in terms of the classes that achieved good level of discrimination from others in the training set.

Firstly, our proposed system will reduced some features in training dataset by using Random Forest algorithm to each connection at filtering in preprocessing state. The system will use information Gain Algorithm to eliminate as Fig. 2.

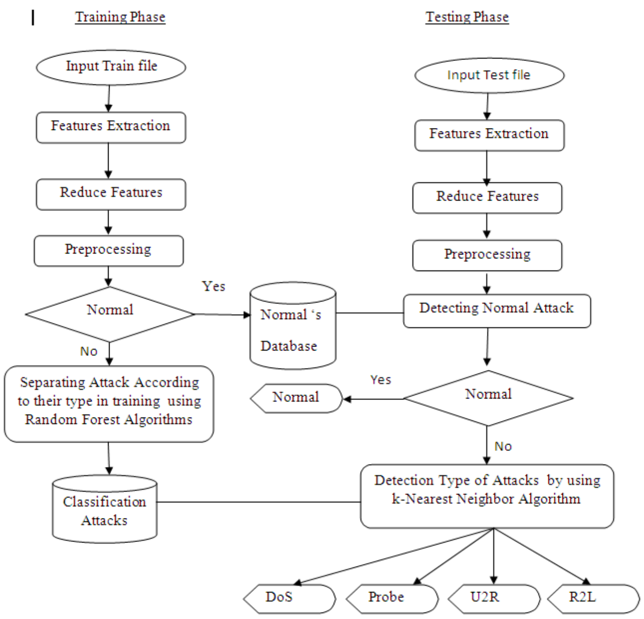
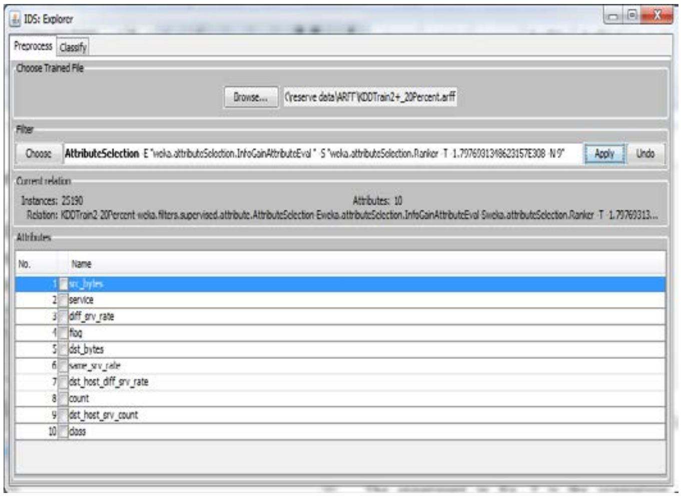


Fig. 1 Proposed System



C:\reserve data\ARIT\KDDTrain2+_20Percent.arff

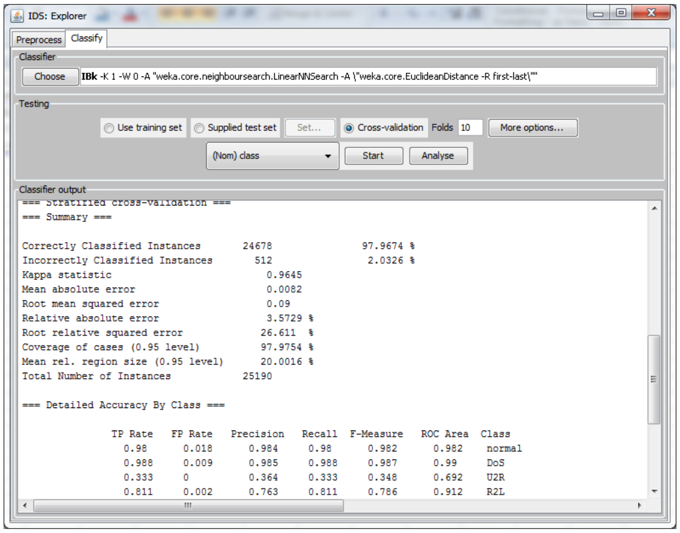
AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 9"

Instances: 25190 Attributes: 10

Relation: KDDTrain2-20Percent-weka.filters.supervised.attribute.AttributeSelection-Eweka.attributeSelection.InfoGainAttributeEval-Sweka.attributeSelection.Ranker -T -1.79769313...

No.	Name
1	src_bytes
2	service
3	diff_srv_rate
4	flag
5	dst_bytes
6	same_srv_rate
7	dst_host_diff_srv_rate
8	count
9	dst_host_srv_count
10	class

Fig. 2 Ranking and filtering mode evaluated by Random Forest for feature selection in proposed system



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set ○ Supplied test set ● Cross-validation Folds 10

(Nom) class

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       24678               97.9674 %
Incorrectly Classified Instances       512                2.0326 %
Kappa statistic                          0.9645
Mean absolute error                      0.0082
Root mean squared error                  0.09
Relative absolute error                  3.5729 %
Root relative squared error             26.611  %
Coverage of cases (0.95 level)          97.9754 %
Mean rel. region size (0.95 level)      20.0016 %
Total Number of Instances            25190

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.98      0.018      0.984     0.98      0.982      0.982    normal
                 0.988     0.009      0.985     0.988     0.987      0.99     DoS
                 0.333     0          0.364     0.333     0.348      0.692    U2R
                 0.811     0.002      0.763     0.811     0.786      0.912    R2L
```

Fig. 3 Classification step using with KNN in proposed system

So, system will try to detect various anomaly attacks using KDD dataset with KNN classification algorithm. The proposed system will reduced in training time and will increase the accuracy of the system's classification, also will solve imbalance intrusions problem by increasing the accuracy on minority attacks.

For the improvement of detection rate on our proposed system, we examined the number of features to use by ranking and reducing features and getting the best experimental results as shown in fig (3), getting the best number of features is 9.

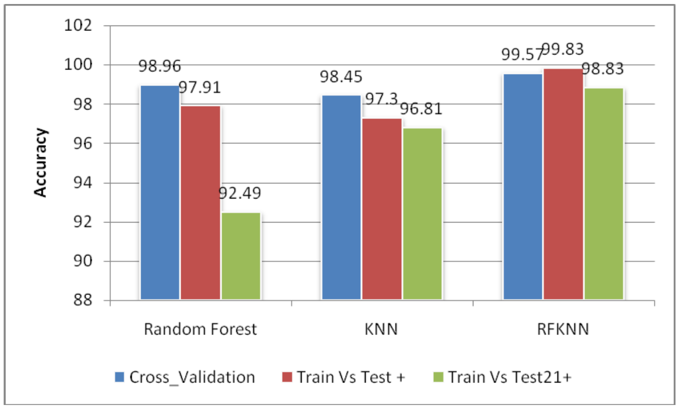


Fig. 3 The detection rates between proposed method RFKNN with RF and KNN in overall attacks type

In the experiments process, the system use 10 trees and reduced 9 features in my proposed method to classify. The accuracy of the system will be increased other systems in overall and the detection rate using proposed method on each minority attack types.

The experiment result in fig. 3 is the comparison between Random Forest, KNN and the proposed method by classifying with training dataset in cross-validation mode and testing with test datasets. The results will shown that the proposed method can detect in more precisely than other.

http://dx.doi.org/10.15242/IIE.E0314114

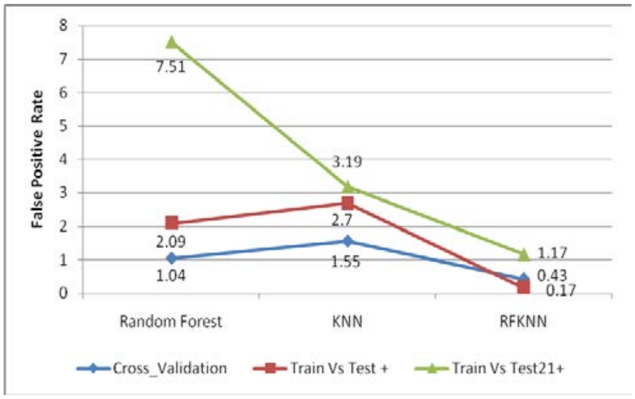


Fig. 3 The false positive rates between proposed method RFKNN with RF and KNN in overall attacks type

As a results shown in Fig. 4 , the proposed method will reduce the false positive rate more than each separated algorithm significantly.

V.CONCLUSION

Many researchers are presented as a classical tool for Intrusion Detection System as separated algorithm. But in each algorithm has the weakness and draw backs in timing and performance. So, we intend to solve this by combining the advantages of feature selection of Random Forest and the pattern classification of KNN. System can detected more precisely and reduce in false positive rate. These experiments prove that the proposed method can stand as a classical tool for IDSs. By using this proposed method, data can be reduced in size and process time. And the accuracy can get more.

ACKNOWLEDGMENT

I would like to express my respectful gratitude to our Rector of University of Technology (Yatanarpon Cyber City),Dr. Aung Win and Department head, Faculty of Information and Communication Technology, Dr. Soe Soe Khaing, for allowing me to develop my research and giving me general guidance during period of my study.

REFERENCES

[1] G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64. F. Anjum, D. Subhadrabandhu and S. Sarkar, “Signature Intrusion Detectio for Wireless Ad Hoc Networks: A Comparative study of various routing protocols”, in 2003.

[2] P C Kishore Raja, Dr.Suganthi.M, R.Sunder, “wireless node behavior based intrusion detection using genetic algorithm”, Ubiquitous Computing and Communication Journal, 2006.

[3] Y. Chihab, A. A. Ouhman, M. Erritali and B. E. Ouahidi, “Detection & Classification of Internet Intrusion Based on the Combination of Random Forest and Naïve Bayes,” Internal Journal of Engineering and Technology, IJET, Vol 5 No 3 Jun-Jul 2013. ISSN : 0975-4024 p. 2116-2126

[4] J. Ross Quinlan, Machine Leaming, 1986, “Induction of decision trees,” p. 81-106.

[5] J. Zhang and M. Zulkernine,” Network Intrusion Detection using Random Forests,” School of Computing Queen’s University, Kingston Ontario, Canada.

[6] D. J. Hand, H. Mannila and P. Smyth, Principles of Data Mining, The MIT Press,

[7] L. Breiman, “Random Forests”, Machine Learning 45(1):5–32, 2001. http://dx.doi.org/10.1023/A:1010933404324

[8] J. Zhang and M. Zulkernine, ”Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection”, Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul,Turkey, June 2006.

[9] T. Lappas and K. Pelechrinis, “Data mining Techniques for(Network)Intrusion Detection System,”, UC Riverside, Riverside CA92521.

[10] WEKA software, Machine Learning, http://www.cs.waikato.ac.nz/ml/weka/, The University of Waikato, Hamilton, New Zealand.

[11] M. Dash and H. Liu, “Feature selection for classification,”International Journal of Intelligent Data Analysis, 1(3), 1997.

[12] F. Tan, "Improving Feature Selection Techniques for Machine Learning" (2007). Computer Science Dissertations. Paper 27. http://digitalarchive.gsu.edu/cs_diss/27

[13] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[14] CERT, http://www.cert.org/stats/, 2010.

[15] Khalil El-Khatib “Impact of Feature Reduction on the Efficiency of Wireless Intrusion Detection Systems” IEEE Transactions on Parallel and Distributed Systems, Vol. 21, no. 8, August 2010. http://dx.doi.org/10.1109/TPDS.2009.142